D54–D56 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 18 October 2011*
doi:10.1093/nar/gkr854

The sequence read archive: explosive growth of sequencing data

Yuichi Kodama[1,*], Martin Shumway[2] and Rasko Leinonen[3] on behalf of the International Nucleotide Sequence Database Collaboration

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan, [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 15, 2011; Accepted September 23, 2011

ABSTRACT

New generation sequencing platforms are producing data with significantly higher throughput and lower cost. A portion of this capacity is devoted to individual and community scientific projects. As these projects reach publication, raw sequencing datasets are submitted into the primary next-generation sequence data archive, the Sequence Read Archive (SRA). Archiving experimental data is the key to the progress of reproducible science. The SRA was established as a public repository for next-generation sequence data as a part of the International Nucleotide Sequence Database Collaboration (INSDC). INSDC is composed of the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). The SRA is accessible at www.ncbi.nlm.nih.gov/sra from NCBI, at www.ebi.ac.uk/ena from EBI and at trace.ddbj.nig.ac.jp from DDBJ. In this article, we present the content and structure of the SRA and report on updated metadata structures, submission file formats and supported sequencing platforms. We also briefly outline our various responses to the challenge of explosive data growth.

THE SEQUENCE READ ARCHIVE

Massively parallel next-generation sequencing platforms are revolutionizing life sciences. These instruments are producing vastly more sequence data than that was ever possible with the capillary technology. National Center for Biotechnology Information (NCBI) started the archive of raw sequencing data from next-generation platforms in 2007, followed by European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ) in 2008. In 2009, an international public archival resource 'Sequence Read Archive (SRA)' for next-generation sequencing data was established as a part of the International Nucleotide Sequence Database Collaboration (INSDC) (1–3). The mission of the SRA is to help the wider research community gain access to the next generation sequencing data emanating from scientific research. The SRA works as a core infrastructure for sharing of pre-publication sequence data as required by several large-scale international projects including the Human Microbiome project (https://commonfund.nih.gov/hmp) and 1000 Genomes project (http://www.1000genomes.org). It is to be noted that data requiring authorized access, such as human genome sequenced under ethical consent agreements, should be submitted to the database of phenotypes and genotypes at NCBI (dbGaP, http://www.ncbi.nlm.nih.gov/gap) or to the European Genome-phenome Archive at EBI (EGA, http://www.ebi.ac.uk/ega). Data submitted to dbGaP or EGA is not part of the public SRA. However, summary-level metadata is made available through SRA.

CONTENT

In 2011 the SRA surpassed 100 Terabases of open-access genetic sequence reads from next generation sequencing technologies. The Illumina™ platform comprises 84% of sequenced bases, with SOLiD™ and Roche/454™ platforms accounting for 12% and 2%, respectively. The most active SRA submitters in terms of submitted bases are the Broad Institute, the Wellcome Trust Sanger Institute and Baylor College of Medicine with 31, 13 and 11%, respectively. The largest individual global project generating next-generation sequence is the 1000 Genomes project which has contributed nearly one third of all bases. The most sequenced organisms are *Homo sapiens* with 61%, human metagenome with 6% and *Mus musculus* with 5% share of all bases. The common

*To whom correspondence should be addressed. Tel: +81 55 981 6853; Fax: +81 55 981 6849; Email: ykodama@genes.nig.ac.jp

study types in terms of sequenced bases are Whole Genome Sequencing and Re-sequencing, Population Genomics, Metagenomics and Epigenetics with 57, 12, 11 and 8% share of all bases, respectively.

ACCEPTED DATA

The SRA is a repository of raw sequence data with the aim to balance the cost of long-term archival with the requirement to store sufficient information to support re-use of the submitted data. At minimum, data submitted to SRA must include base or SOLiD color calls and their qualities. To limit the archival cost and guided by community consultation, the SRA also sets maximum levels for accepted raw data. For example, since the end of 2010 signal data from the Illumina and SOLiD platforms are no longer archived by the SRA. In addition to base calls (or SOLiD color calls) and quality scores, SRA also accepts alignments submissions in BAM (4) format. Other data may be accepted as well; full details are available for submitters from NCBI, DDBJ or EBI. Interactive and pipeline submission routes to the SRA archives are available. Functional genomics studies using next-generation sequencing (e.g. ChIP-seq and RNA-seq) can be submitted via the Gene Expression Omnibus at NCBI (http://www.ncbi.nlm.nih.gov/geo) (5), ArrayExpress at EBI (http://www.ebi.ac.uk/arrayexpress) (6) and DDBJ Omics Archive (http://trace.ddbj.nig.ac.jp/dor) (7).

SUPPORTED PLATFORMS AND FILE FORMATS

The SRA aims to support all established and emerging sequencing platforms and most commonly used data file formats. Supported platforms include Roche/454 (Roche Diagnostics Corp.), Illumina (Illumina Inc.), SOLiD (Life Technologies Corp.), HeliScope™ Single Molecule Sequencer (Helicos Biosciences Corp.), Complete Genomics™ (Complete Genomics Inc.), SMRT™ (Pacific Biosciences Inc.) and Ion Torrent PGM™ (Life Technologies Corp.). Depending on the data file format, submissions for emerging platforms may be first supported only provisionally where submitted data is made available only in the original submitted format. This procedure guarantees early access to data generated by new platforms. Data submitted in any of the widely used data formats is rigorously validated and made available to pubic in a variety of formats. For example, NCBI makes data available in the NCBI SRA toolkit format, which can be converted into many other file formats, while EBI and DDBJ make data available in the FASTQ format. Recommended data submission formats may vary slightly between DDBJ, EBI and NCBI, but all widely used formats, such as BAM and Standard Flowgram Format (SFF), are universally accepted.

METADATA MODEL

Data submitted to SRA is organized using a metadata model consisting of six objects: study, sample, experiment, run, analysis and submission. The SRA study contains high-level information including goals of the study and literature references, and may be linked to the INSDC BioProject database. Similarly, the SRA sample object contains detailed sample information, and may be linked to the BioSample databases of NCBI (http://www.ncbi.nlm.nih.gov/biosample) and EBI (http://www.ebi.ac.uk/biosamples). The SRA experiment and run objects contain library and instrument information and are directly associated with the sequence data. The SRA analysis object is used for the deposition of a variety of analysis results including alignments and assemblies. The SRA submission object groups the other objects for submission into the SRA. These metadata objects are all accessioned with unique permanent identifiers that are shared by INSDC partners.

The SRA has updated the metadata model to better represent new sequencing technologies and applications. The schema version 1.3 introduced in 2011 added a new structure called GapDescriptor that will encode the placement of spot subsequences (tags) against a reference or assembly substrate. This structure encodes mate pair gaps and tandem read gaps. Introduction of the GapDescriptor element was motivated by the need to describe Complete Genomics platform sequencing. The next planned metadata version, 2.0, will largely simplify the model by removing redundant and deprecated fields. While this new model will be incompatible with the previous version, the SRA archives will transform all existing metadata documents to conform to the new model. The SRA metadata model is largely shared by all three archives, however, small differences have been introduced to support archive specific local requirements.

SEQUENCE DATA EXCHANGE

The public sequence data are exchanged between the INSDC partners allowing all public data to be accessed at each site regardless of the point of the original submission ('submit locally, share globally' model). The data is currently exchanged in the NCBI SRA toolkit format (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). The SRA toolkit provides a configurable storage and compression architecture and its format can be converted to other formats such as widely-used FASTQ through its standard API. The SRA data exchange model follows the long established INSDC policy of exchanging GenBank, EMBL-Bank and DDBJ entries.

CHALLENGE OF DATA GROWTH

The explosive growth of next-generation sequencing data submitted into the SRA exceeds the growth rate of storage capacity. This trend provides the greatest challenge to handle raw sequence data for SRA archives and users of the raw sequence data alike. The SRA partners actively discuss and pursue approaches together with user communities to maximize the benefit gained from archiving next-generation sequencing data while minimizing the infrastructure costs. Possible approaches discussed

include reference-based compression of sequencing data, quantization of base quality values, selective storage of base quality values, reducing the metadata stored for individual reads (e.g. read names), federation of data in place of data submission and exchange, and consolidation of catastrophe back-up storage across SRA archives. Among these possibilities, SRA is exploring approaches based on reference alignment and compression of reads, and on the preservation of only the most valuable base quality information (8), and is also actively participating in experiments assessing the effect of quality score quantization. The SRA partners continue actively to discuss with the research community to explore appropriate data reduction approaches.

FUNDING

DNA Data Bank of Japan, Ministry of Education, Culture, Sports, Science and Technology of Japan; European Molecular Biology Laboratory, European Commission and the Wellcome Trust; National Library of Medicine; Intramural Research Program of the NIH. Funding for open access charge: Ministry of Education, Culture, Sports, Science and Technology of Japan (management expense grant).

Conflict of interest statement. None declared.

REFERENCES

1. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
2. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
3. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. (2012) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*, **40**, D33–D37.
4. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics.*, **25**, 2078–2079.
5. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
6. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
7. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. The DNA Data Bank of Japan launches a new resource DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, in press.
8. Fritz,M.H.-Y., Leinonen,R., Cochrane,G. and Birney,E. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, **21**, 734–740.